

1.1 Brief history of Bioinformatics

In 1866, Gregor Johann Mendel discovered genetics. Mendel's hybridization experiments on peas unveiled some biological elements called genes, which pass information from generation to generation. At that time, people thought genetic information was carried by some "chromosomal protein", however it is not.

Later in 1869, DNA was discovered. But it was not until 1944 that Avery and McCarty demonstrated DNA is the major carrier of genetic information, but not protein. Remarkable as it is, this discovery is often referred as the start of bioinformatics. In 1953, another historic discovery enabled great advances in both biology and bioinformatics: James Watson and Francis Crick deduced the three dimensional structure of DNA, which is a double helix.

Later in 1961, the genetic code, namely how the mapping from DNA to peptide (protein) is done, was elucidated by Marshall Nirenberg. It is by the means of combining three nucleotides in the DNA as a codon, and mapping each of them to one amino acid in the peptide.

In 1968, the restriction enzyme was discovered and isolated from bacteria. These enzymes protect the bacteria by cutting any foreign DNA molecules at specific sites so as to restrict the ability of the foreign DNA molecules to take over the transcription and translation machinery of the bacterial cell.

Starting from the 1970's, several important biotechnology techniques were developed. Firstly, DNA sequencing techniques, like sequence segmentation and electrophoresis were developed. These enabled the identification of DNAs given just a tissue found on a human-body. Moreover, in 1985, the groundbreaking technique, Polymerase-Chain-Reaction (PCR) was invented. By exploiting natural replication, DNA samples can be easily amplified using PCR, so that they are enough for doing experiment.

In 1986, RNA splicing in eukaryotes was discovered. This is the process of removing introns and rejoining the exons in order to produce a functional mRNA from a pre-mRNA. The splicing is carried out by spliceosomes, which is assembled from a group of smaller RNA-protein complexes known as snRNPs and additional proteins.

Starting from the 1980's, scientists began to sequence the genomes. From 1980-1990, complete sequencing of the genomes of various organisms, like that

of the E.Coli, was done successfully. And probably the most remarkable event was the launch of Human Genome Project (HGP) in 1989. Originally, it was planned to be completed in 15 years; however, thanks to more and more emerging advanced techniques, in the year of 2001, the first draft of the human genome was published. Subsequently, a more refined human genome was also published in 2002.

Triggered by the Human Genome Project, the Genomes to Life(GTL) project was launched in 2001 and aimed to complete in the next 10 to 20 years. The objectives of the project was to understand the detailed mechanism of cells, for instance, identifying the proteins involves in sustaining critical life functions and characterizing the gene regulatory networks that regulate the expression of those proteins. The diverse abilities of complex microbial communities will also be explored in addressing DOE missions. Lastly, these data will be integrated and information are extracted as much as possible by computational tools and modelling.

1.2 Cell

Our body consists of a number of organs. Each organ composes of a number of tissues, and each tissue composes of cells of the same type. The individual cell is the minimal self-reproducing unit in all living species. It performs two type of functions, i.e. performs chemical reactions necessary to maintain our life and also passes the information for maintaining life to the next generation. Since the cell is the vehicle for transmission of the genetic information in all living species, it needs to store the genetic information in the form of double-stranded DNA. The cell replicates its information by separating the paired DNA strands and using each as a template for polymerization to make a new DNA strand with a complementary sequence of nucleotides. The same strategy is used to transcribe portions of the information from DNA into molecules of the closely related polymer, RNA. RNA is the intermediate between DNA and protein and it guides the synthesis of protein molecules by the complex machinery of translation, i.e. the ribosome. The resultant proteins are the main catalysts for almost all the chemical reactions in the cell. In addition to catalyst, proteins are performing also building block, transportation, signalling, etc.

1.3 DNA, RNA, Protein

1.3.1 Protein

Proteins constitute most of a cell's dry mass. They are not only the building blocks from which cells are built; they also execute nearly all cell functions.

Understanding of proteins can guide us to understand how our bodies function and other biological processes.

Protein is made from a long chain of amino acids, each links to its neighbor through a covalent peptide bond. There are 20 types amino acids in proteins, and each amino acid carries different chemical properties. The length of protein is in the range of 20 to more than 5000 amino acids. In average, protein contains around 350 amino acids. Therefore, protein is also known as polypeptides.

In order to perform its chemical function, proteins need to fold into certain 3 dimensional shapes. There are several interactions that cause the proteins to fold, such as the sets of weak noncovalent bonds that form between one part of the chain and another. The weak bonds are of three types, i.e. hydrogen bonds, ionic bonds, and van der Waals attractions. In addition to these three weak bonds, the fourth weak force, i.e. the hydrophobic interaction also has a central role in determining the shape of a protein. Correct shape for a protein is vital to its functionality.

1.3.1.1 Amino Acids

As we have known, amino acid is the building block of proteins. Understanding the properties of amino acids will help us in understanding the folding and function of proteins.

Now, let's look at the composition of amino acid. Each amino acid consists of:

1. Amino Group (-NH₂ group)
2. Carboxyl Group (-COOH group)
3. R Group (Side Chain), which determine the type of amino acid

All the groups are attached to a single carbon atom called α -carbon.(see Figure 1.1)

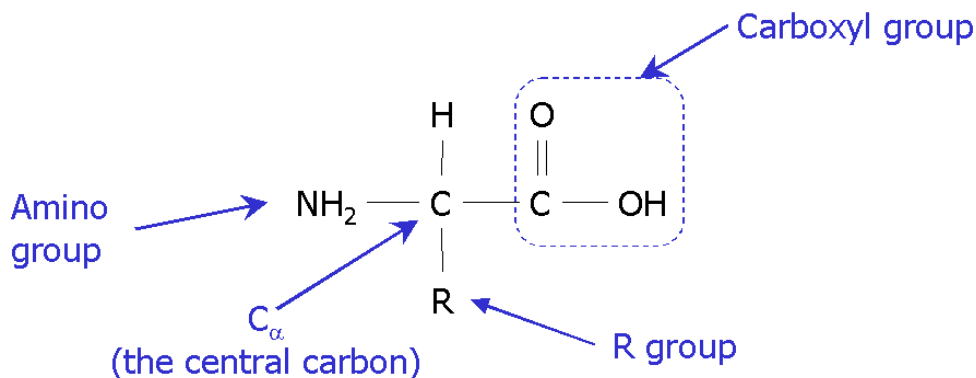


Figure 1.1: Structure of Amino Acid

Amino acids can be classified into 4 types, i.e.:

1. Positively charged (basic) amino acids include Arginine (Arg, R), Histidine (His, H), Lysine (Lys, K)
2. Negatively charged (acidic) amino acids include Aspartic acid (Asp, D), Glutamic acid (Glu, E)
3. Polar amino acids include Asparagine (Asn, N), Cysteine (Cys, C), Glutamine (Gln, Q), Glycine (Gly, G), Serine (Ser, S), Threonine (Thr, T), Tyrosine (Tyr, Y). Polar amino acids are overall uncharged, but they have uneven charge distribution. They can form hydrogen bonds with water so they are called hydrophilic amino acids. They are often found on the outer surface of a folded protein.
4. Nonpolar amino acids include Alanine (Ala, A), Isoleucine (Ile, I), Leucine (Leu, L), Methionine (Met, M), Phenylalanine (Phe, F), Proline (Pro, P), Tryptophan (Trp, W), Valine (Val, V). Nonpolar amino acids are overall uncharged and have uniform charge distribution. They cannot form hydrogen bonds with water and therefore they tend to appear on the inside surface of a folded protein. They are called hydrophobic amino acids.

As we have known, protein or polypeptide chain is formed by joining the amino acids together via a peptide bond. The formation and the structure of the peptide bond are shown in Figure 1.2.

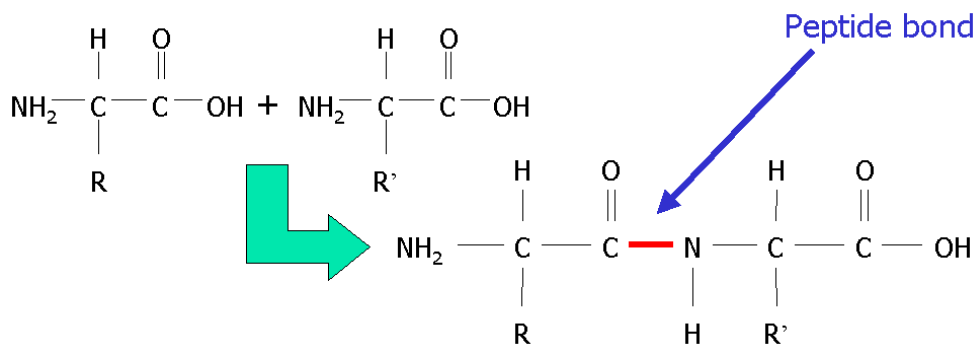


Figure 1.2: Formation and Structure of the Peptide Bond

One end of the polypeptide is the amino group, which is called N-terminus. The other end of the polypeptide is the carboxyl group which is called C-terminus.

1.3.2 DNA

DNA is the genetic material in all organisms (with certain viruses being exception) and it stores the instruction needed by the cell to perform daily life function.

DNA can be thought of as a large cookbook with recipes for making every protein in the cell. The information in DNA is used like a library. Library books

can be read and reread many times. Similarly, the information in the genes is read, perhaps millions of times in the life of an organism, but the DNA itself is never used up.

DNA consists of two strands which interwoven together and form a double helix. Each strand is a chain of small molecules called nucleotides.

1.3.2.1 Nucleotides

Nucleotides are the building blocks of all nucleic acid molecules (such as DNA and RNA). These structural units consist of three essential components, i.e.

1. A pentose sugar – deoxyribose (in DNA) and Ribose (in RNA)
2. Phosphate (bound to the 5' carbon)
3. Base (bound to the 1' carbon) – nitrogenous base

For the structure of nucleotides (in DNA), please refer to Figure 1.3.

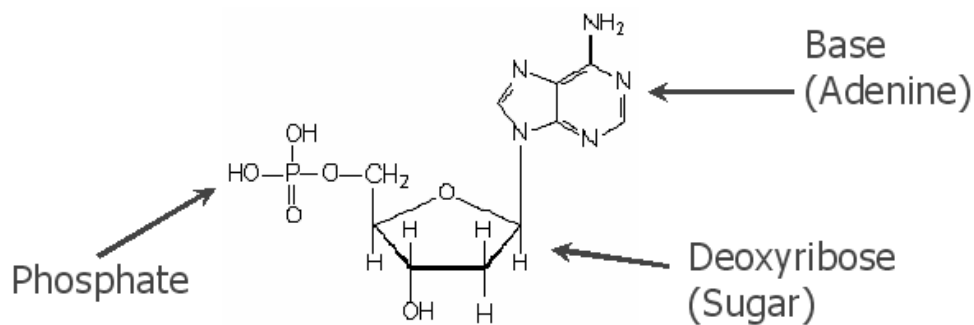


Figure 1.3: Structure of Nucleotide that forms DNA

The nucleotide structure can be broken down into 2 parts, i.e. the sugar-phosphate backbone and the base. All nucleotides share the sugar-phosphate backbone. Nucleotide polymers are formed by linking the monomer units together using an oxygen on the phosphate, and a hydroxyl group on the sugar. Joining of two nucleotides forms a dinucleotide; of three nucleotides, a trinucleotide; and so forth. Short chains consisting of fewer than 20 nucleotides linked together are called oligonucleotides, and longer chains are referred as polynucleotides.

The pentose sugar found in nucleic acids give them their names. In Ribonucleic Acids (RNA), the pentose sugar is called ribose; while the pentose sugar in Deoxyribonucleic acids (DNA) is called deoxyribose because the OH group in position 2' has been changed to H (i.e. the oxygen has been removed).

There are two kinds of nitrogenous bases, i.e. the nine-membered double-ringed purines and the six-membered single-ringed pyrimidines. There are two types of purines, i.e. Adenine(A) and Guanine(G). And there are three types of pyrimidines, i.e. Thymine(T), Cytosine(C), and Uracil(U). These bases will be explained in more details in the next section.

1.3.2.2 Forms of Nucleotides

Nucleotides can have 1, 2, or 3 phosphate groups. Monophosphate nucleotides have only 1 phosphate, which are the building blocks of DNA. Diphosphate nucleotides have 2 phosphate groups and triphosphate nucleotides have 3 phosphate groups, which are used to transport energy in the cell. Triphosphate nucleotides are shown in the Figure 1.4.

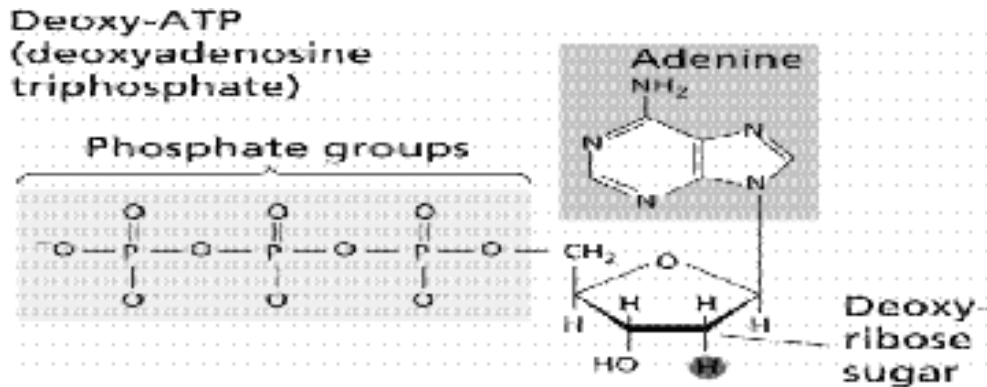


Figure 1.4: Triphosphate Nucleotides

There are five different types of nucleotides, differing only in the nitrogenous base. The five nucleotides are given one letter abbreviations as shorthand for the five bases. Their structures are shown in Figure 1.5.

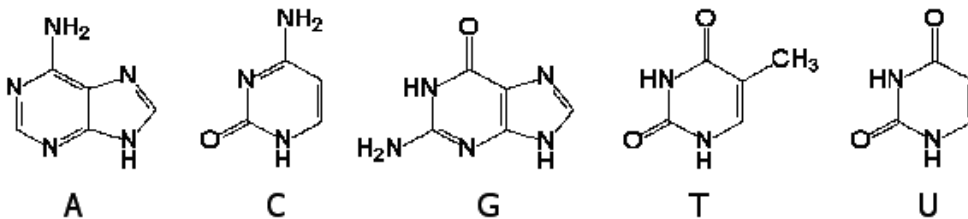


Figure 1.5: Types of Nucleotides, A=Adenine, G=Guanine, C=Cytosine, T=Thymine, U=Uracil

The 3D structure of these nucleotides are shown in Figure 1.6 and 1.7.

The base on each nucleotide is different, but they still show similarities. A and G are called purines, they have two ring structure. C, T and U are pyrimidines, they have one ring structure. Pyrimidine can only base-pair with Purine, just like A base-pair with T or U, and G base-pair with C. A cannot base-pair with G and T cannot base-pair with U or C. For purines, the 9 atoms that make up the fused rings (5 carbon, 4 nitrogen) are numbered 1-9. All ring atoms lie on the same plane. For pyrimidines, the 6 atoms (4 carbon, 2 nitrogen) are numbered 1-6. Like purines, all pyrimidine ring atoms lie in the same plane. The nucleotides for DNA is A, G, C, and T. Whereas the nucleotides for RNA is A, G, C, and U.

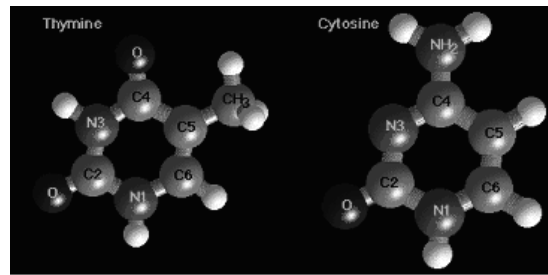


Figure 1.6: 3D view of Nucleotides – Cytosine and Thymine

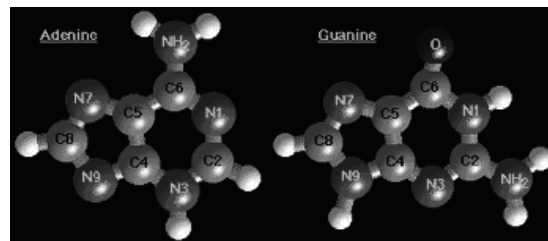


Figure 1.7: 3D view of Nucleotides – Adenine and Guanine

1.3.2.3 DNA Structure

DNA is double-helix in structure and it consists of two strands which interwoven together to resemble a twisted ladder. (see Figure 1.8).

If you look at it in detail, you could observe that the rungs are consisted of chemical compounds called bases, while the sides of the rungs are the sugar (deoxyribose) and the phosphate molecules. These three parts, i.e. base, sugar, and phosphate form the small molecules that we knew as nucleotides. There are 4 types of bases that form the rungs of DNA double-helix, i.e. the 4 letters genetic code (A/Adenine, G/Guanine, C/Cytosine, and T/Thymine).

The correct structure of DNA was first deduced by J. D. Watson and F. H. C. Crick in 1953. Their double helix model of DNA structure was based on two major kinds of evidence:

1. From the analysis of E. Chargaff and colleagues, the concentration of Thymine was always equal to the concentration of Adenine and the concentration of Cytosine was always equal to the concentration of Guanine. This observation strongly suggests that A and T as well as C and G have some fixed relationship.
2. X-Ray diffraction pattern from R. Franklin, M. H. F. Wilkins and coworkers. The data indicated that DNA has a highly ordered, multiple-stranded structure with repeating substructures spaced every 3.4 angstroms along the axis of the molecule.

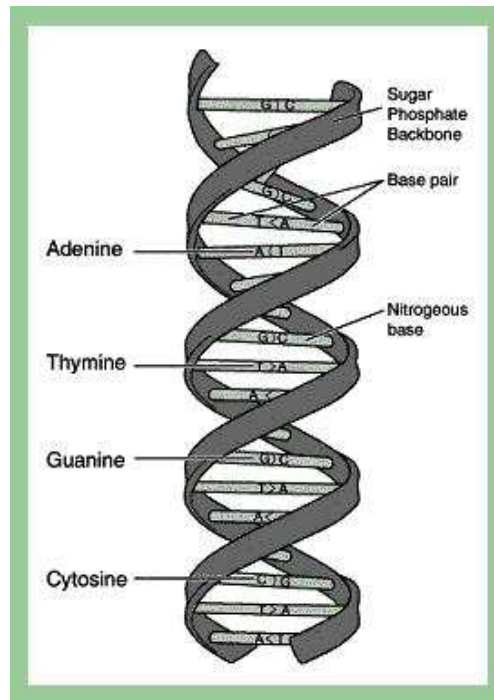


Figure 1.8: Structure of DNA

From these studies, Watson and Crick proposed that DNA exists as a double helix in which polynucleotide chains consist of a sequence of nucleotides linked together by phosphodiester bonds, joining adjacent deoxyribose moieties. The two polynucleotide strands are held together by the hydrogen bonding between bases in opposing strands. The base-pairs are of high specificity such that: A is always paired with T and G is always paired with C. These base pairs are called as the complementary base-pairing. A can form 2 hydrogen-bonds with T, whereas G can form three hydrogen-bonds with C. Please look at Figure 1.9 for the base-pairing.

Since A only pairs with T and G only pairs with C, the double-stranded DNA (if unwind) would look like Figure 1.10)

The reasons behind the complementary bases are:

1. Purines (A or G) cannot pair up because they are too big
2. Pyrimidines (C or T) cannot pair up because they are too small
3. G and T (or A and C) cannot pair up because they are chemically incompatible.

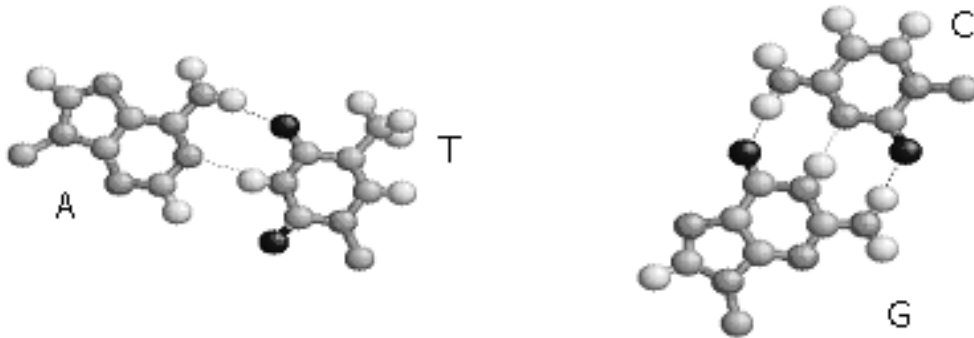


Figure 1.9: Watson-Crick Base-Pairing

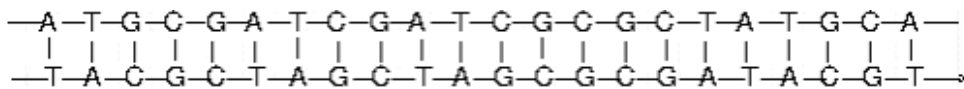


Figure 1.10: Example of double strand DNA, the two strands show a complementary base-pairing

1.3.2.4 Orientation of a DNA

One strand of DNA is generated by chaining together nucleotides. It forms a phosphate-sugar backbone. It has direction: from 5' to 3' (upstream). The complementary strand goes from 3' to 5' (downstream). DNA always extends from 3' end. The strand from 3' to 5' is the reverse complement of the strand from 5' to 3'. Reverse complement means that if the base in the 5' to 3' strand is A, then the base in the 3' to 5' strand is T. The reverse complement follows the Watson-Crick base-pairing (A is to T and G is to C). The orientation of DNA is shown in Figure 1.11.

The DNA exists as a double-stranded molecule within a cell. They are antiparallel; that is they have opposite chemical polarity. As one moves unidirectionally along a DNA double helix, the phosphodiester bonds in one strand go from a 3' carbon of one nucleotide to a 5' carbon of the adjacent nucleotide, while those in the complementary strand go from a 5' carbon to a 3' carbon. This opposite polarity of the complementary strands is very important in considering the mechanism of replication of DNA.

1.3.2.5 Circular Form of DNA

DNA usually exists in linear form, e.g. the DNA in human and yeast exists in linear form. However, in some simple organism, DNA exists in circular form, e.g. Prokaryotes, Mitochondria, Chloroplasts, Viruses. The circular form of DNA is shown in Figure 1.12.

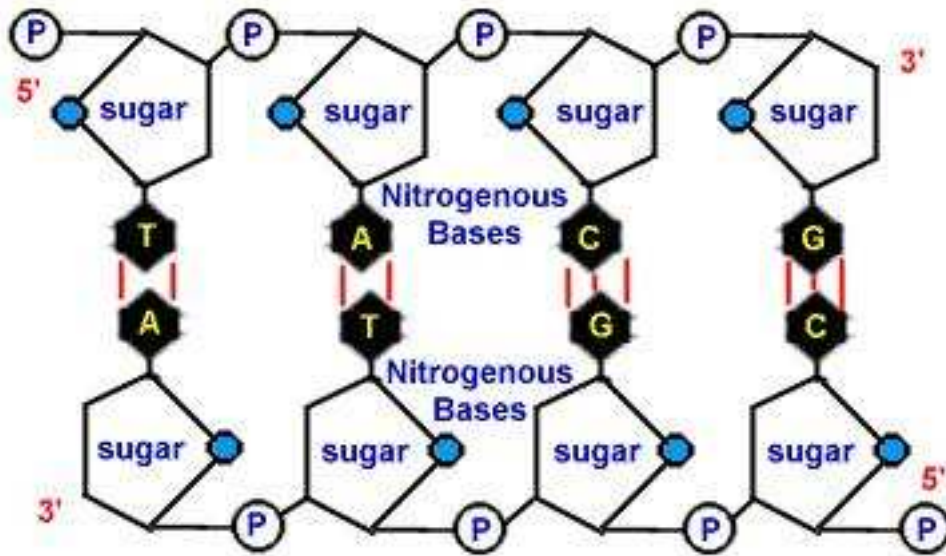


Figure 1.11: Orientation of DNA



Figure 1.12: Circular DNA

1.3.2.6 Location of DNA in a Cell

There are two types of organisms, i.e. Prokaryotes and Eukaryotes. Prokaryotes are single-celled organisms with no nuclei (e.g. bacteria). They have no distinct nuclear compartment to house their DNA and therefore the DNA swims within the cells. Eukaryotes, on the other hand, are organisms whose cells contain a nucleus surrounded by cytoplasm which is contained within a plasma membrane. The DNA locates within the nucleus. Eukaryotes are organisms with single or multiple cells, for example, plant and animal.

Please refer to Figure 1.13 for Prokaryote cells and Figure 1.14 for Eukaryote cells.

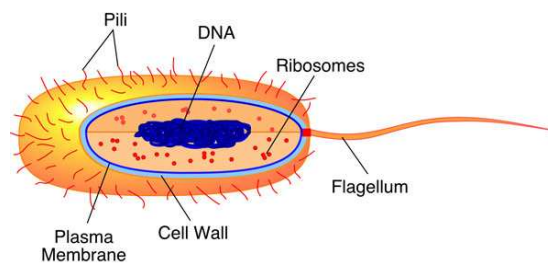


Figure 1.13: Typical Prokaryotic Cell

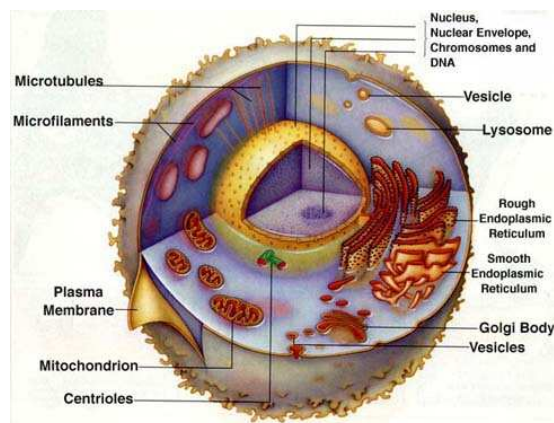


Figure 1.14: Typical Eukaryotic Cell

1.3.3 RNA

We have discussed both DNA and protein, and now we will go on to RNA. RNA is the nucleic acid which is produced during the transcription process (i.e. from DNA to RNA). However, in certain organisms, such as viruses, they carry RNA

as their genetic material instead of DNA.

There are 3 major types of RNA, i.e. messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Messenger RNAs carry the encoded information required to make proteins of all types. Ribosomal RNAs form parts of ribosomes that helps to translate mRNAs into proteins. Transfer RNAs serve as a molecular dictionary that translates the nucleic acid code into the amino acid sequences of proteins.

As we have mentioned before, RNA has the properties of both DNA and protein. First, similar to DNA, it can store and transfer information. Secondly, similar to protein, it can form complex 3D structure and perform some functions. So, it seems that we only need RNA to accomplish all the requirements for DNA and protein. Why we still need DNA and protein? The reason lies in a simple rule - when you want to do two different things at the same time, you can never do either one as perfectly as those people that only focus on only one thing. As the storage of information, RNA is not as stable as DNA, and that's why we still have DNA. And protein can perform more functions than RNA do, which is the reason that we still need protein.

1.3.3.1 Nucleotide for RNA

We have illustrated the nucleotide structure of DNA before. Figure 1.15 shows the nucleotide structure for RNA. Similar to the nucleotide of DNA, the nucleotide for RNA also has Phosphate and Base. The only difference is that the nucleotide here has Ribose Sugar, instead of Deoxyribose in the DNA nucleotide. The Ribose has an extra OH group at 2', which is different from the H group at the same place of Deoxyribose. That's why we call these two different things "Ribonucleic Acid" and "Deoxyribonucleic Acid" - one is with the OH group, which contains the "O" molecule, yet the other one without.

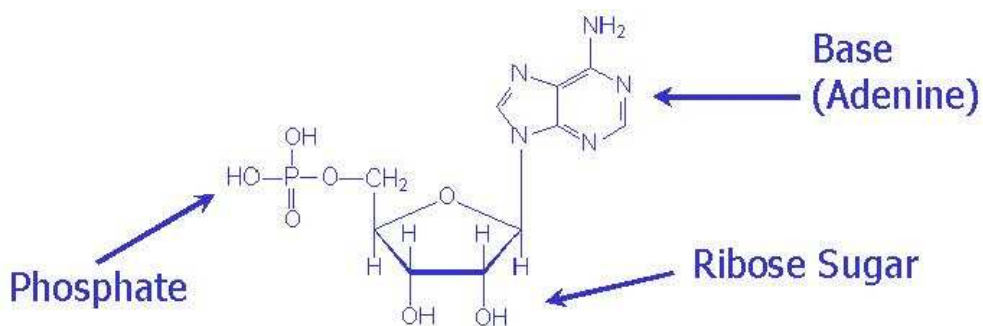


Figure 1.15: Diagram of nucleotide for RNA

1.3.3.2 RNA vs DNA

Besides the primitive difference that one OH group takes place of the H group, RNA has some other characteristics so that we can easily differentiate it from DNA. First of all, unlike the double helix structure of DNA, RNA is single-stranded. One might doubt that with the simple single strand structure, RNA should perform even fewer functions than DNA. The hint here is just the extra OH group. Due to this extra OH, RNA can form more hydrogen bonds than DNA, so that it can form complex 3D structure to perform more functions. And finally, RNA uses Base U instead of the Base T that DNA uses. Base U is chemically similar to Base T. In particular, U is also complementary to A.

1.4 Genome, Chromosome, and Gene

1.4.1 Genome

The *genome* of an organism is its complete set of DNA. All the genetic information in an organism is referred collectively as a “genome”. Genomes vary widely in size: the smallest known genome for a free-living organism (a bacterium of the genus *Mycoplasma*, such as *Mycoplasma genitalium*) contains about 600,000 DNA base pairs, while human and mouse genomes have some 3 billions. Except for mature red blood cells, all human cells contain a complete genome. In most multi-cell organisms, every cell contains the same complete set of genome, except some may have small difference due to mutation. However, in the sperm and egg cell, they only contain half of the genome. This is due to the fact of maintaining the correct size of genome in an individual organism. During fertilization, the sperm cell fuses with the egg cell, and therefore the new cell contains half the genome from the father and half the genome from the mother.

1.4.2 Chromosome

The 3 billion bases of the human genome are not all in one continuous strand of DNA. Rather, the human genome is divided into 23 separate pairs of DNA, called *chromosomes*. Chromosomes are structures within the cell nucleus that carries genes. A chromosome contains a continuous molecule of DNA which is wrapped around histones. Human has 22 pairs of autosomes and 1 pair of sex chromosome, hence make up to 23 pairs of chromosomes. Autosomes are non-sex determining chromosomes, while sex chromosomes are X and Y chromosome. Male will have XY sex chromosomes, whereas female will carry XX sex chromosomes (see Figure 1.16). The collection of chromosomes in an individual is called karyotype. For example, the typical male karyotype has 22 pairs of autosomes, one X and one Y chromosome.

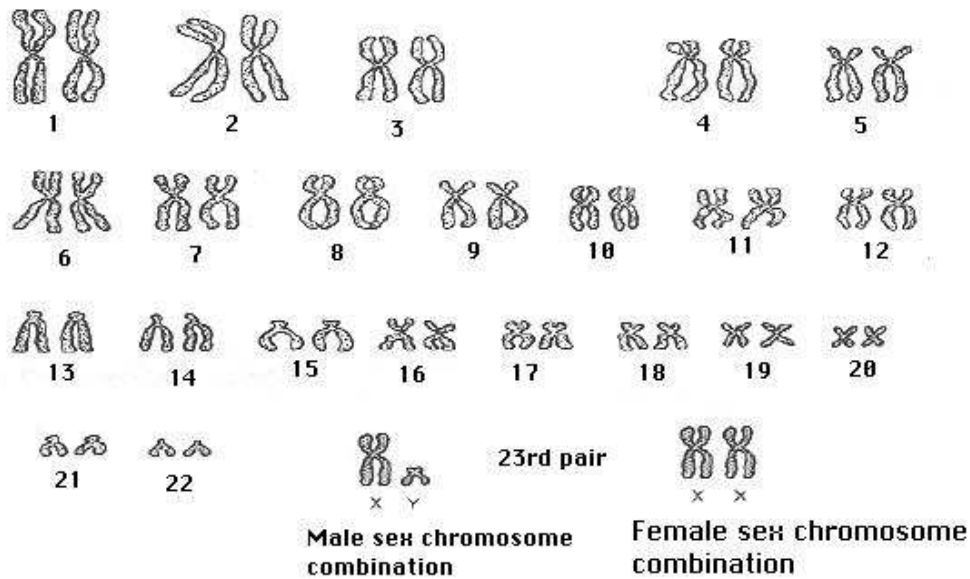


Figure 1.16: Human Chromosomes

1.4.3 Gene

A gene is a DNA sequence that encodes a protein or an RNA molecule. Each chromosome contains many *genes*, i.e. the basic physical and functional units of heredity. Each gene exist in the particular position of particular chromosome. In human genome, it is expected that there are 30,000 - 35,000 genes. In prokaryotic genome, one gene is corresponding to one protein. Whereas, in eukaryotic genome, one gene can corresponds to more than one protein because of the process called as “alternative splicing”. (explain later!)

1.4.4 Complexity of the organism vs. genome size

Due to the huge amount of genes inside human genome, one might argue that the complexity of one organism is somewhat related to its genome size. But in fact, it's not the truth. People have known that the human genome has 3G base pairs, yet the *Amoeba dubia* (a single cell organism), even has up to 670G base pairs! Thus, genome size really has no relationship with the complexity of the organism.

1.4.5 Number of genes vs. genome size

Is there any relationship between the genome size and the number of genes? Before answering this question, let's take a look at the human genome and the genome of another prokaryotic genome, *E. coli*. We have already known that there are about 30,000 to 35,000 genes in human genome, as well as 3G base

pairs. And the biologists have also made the estimation that the average length of the coding region of a gene in human genome is around 1,000 to 2,000 base pairs. So, the total length of all coding regions is less than 70 M base pairs, which is less than 3% of the human genome. For the rest of the genome, people generally call them “junk DNA”.

And how about *E. coli* genome? It has 5M base pairs and 4,000 genes. And the average length of a gene in *E. coli* genome is 1,000 base pairs. From these figures we can know that around 90% of *E. coli* genome is *useful*, i.e., it consists of the coding regions. So, it seems that the Prokaryotic organism, *E. coli*, even has a better genome structure than human beings! In fact, the conclusion here is, for Eukaryotic genome, the genome size has nothing to do with the number of genes!

1.5 Mutation

We have talked about the main usage of DNA, i.e. to transfer information from one generation to another. If such a transfer is always absolutely correct, i.e., the new copy of information is exactly the same as the original one, there would not have any evolution or fatal diseases.

In fact, during the reproduction of DNA, RNA and protein, there exists something called “mutation”. One can take mutation as a sudden change of genome, either in the form of a single gene or in the number of structure of the chromosomes.

For instance, when one segment of DNA is reproduced, a small sub-segment of it could be lost, duplicated or reversed. Furthermore, sometimes a total new segment could be inserted into the DNA segment.

There are several types of mutations. For clarity, please refer to Figure 1.17.

It is the mutation that makes the new generation of cells or organisms might have something different from their ancestor. We can understand that this is just the basis of evolution. But it can also have some evil effect. For example, some “bad” mutation in human genome that alters gene expression may cause fatal disease, such as cancer. Fortunately, mutation occurs in somatic cells is not inherited, only mutation in gametes are passed down. The mutation rate is also typically at 10^{-5} or 10^{-6} . Lastly, mutation can be non-recurrent or recurrent. Non-recurrent mutation is rarely occurred so it is not important. Recurrent mutation occurs frequently at a specific gene locus.

1.6 Central Dogma (from DNA to Protein)

Central Dogma (see Figure 1.18) is the term that tells us how we get the protein from the gene. This process is also called as gene expression. The expression of

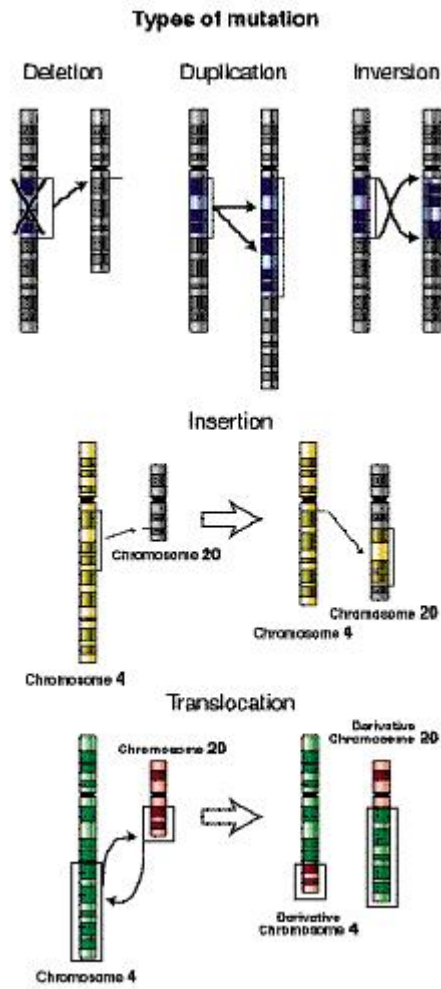


Figure 1.17: Types of mutation

gene consists of two steps, i.e.

1. Transcription: DNA is transcribed to mRNA. During the transcription process, an mRNA is synthesized from a DNA template resulting in the transfer of genetic information from the DNA molecule to the mRNA.
2. Translation: mRNA is translated to Protein. In the translation process, the mRNA is translated to an amino acid sequence by stitching the amino acids one by one during protein synthesis, thus the information obtained from DNA is transferred to the protein.

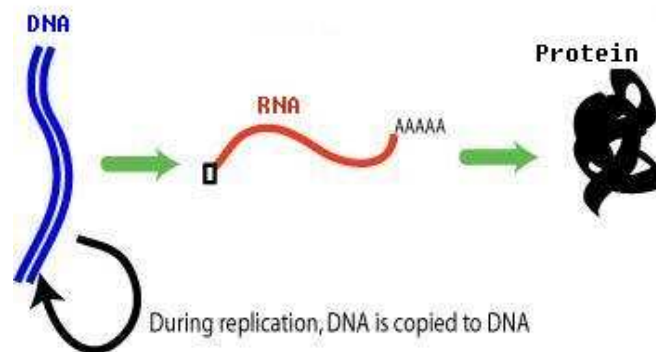


Figure 1.18: Central Dogma: *DNA* is transcribed to *mRNA*, which is transcribed to *Protein*

1.6.1 Central Dogma for Prokaryotes

Please refer to Figure 1.19 for the central dogma of Prokaryotes.

1.6.1.1 Transcription (Prokaryotes)

In general, during the transcription process of prokaryotes, the mRNA is synthesized from one strand of the DNA gene. In this case, RNA polymerase is the enzyme used for transcription. The process of transcription is as follows:

1. An enzyme RNA polymerase temporarily separates the double-stranded DNA
2. It begins the transcription at the transcription start site, which is a kind of marker inside genome.
3. The transcription follows the rule such that the bases A, C and G are copied from DNA to mRNA as exactly the same, but the T is replaced by U (in RNA we only have U instead of T)

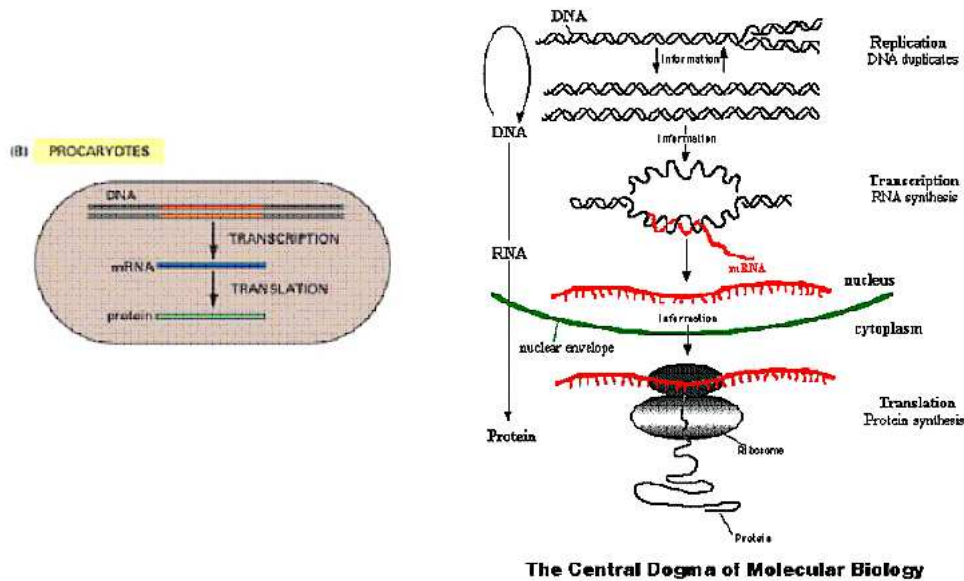


Figure 1.19: Central Dogma for Prokaryotes

- Once the RNA polymerase reaches the transcription stop site (also kinda of marker), the transcription will stop.

1.6.1.2 Translation (Prokaryotes)

Translation process synthesizes a protein from an mRNA. In the translation process, there are three components that are important for protein synthesis, i.e. the Ribosomes (Protein Synthesizing Machines), the tRNA (The adapter molecule), and mRNA (the molecule that carries the genetic information). In the translation process, each amino acid is encoded by a consecutive sequences of 3 nucleotides, known as codon. The decoding table from codon to amino acid is called the "Genetic Code". Since each nucleotide could be one of the four types, A, C, G and U, there are totally $4^3 = 64$ different codons. However, we have already known that there are altogether 20 different amino acids. Thus, the codons are not one-to-one correspondence to the 20 amino acids. From the diagram of the genetic code, we could find out that several different codons code for the same amino acid.(see Figure 1.20). Another important characteristic about the genetic code is that all organisms use the same decoding table. Before we proceed, let's think of why the genetic code is not made of 1 or 2 nucleotides? Recall that there are only 4 types of nucleotides and therefore if genetic code is only made of 1 nucleotide, it can only code for 4 amino acids which is not sufficient. If 2 nucleotides, the number of amino acids can be coded is only $4^2 = 16$ and this is also not sufficient. That's why the genetic code consists of 3 consecutive nucleotides. Since one amino acid is coded from 3 consecutive nucleotides, recall

that the amino acids is only classified into 4 groups and furthermore from the genetic code table, several codons code for same amino acids, therefore a single base change in a codon is usually not sufficient to cause a codon to code for an amino acid from other groups.

1.6.1.3 Genetic code

		Second Position of Codon					
		T	C	A	G		
F i r s t P o s i t i o n	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	T h i r d P o s i t i o n
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA Ter [end]	TGA Ter [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG Ter [end]	TGG Trp [W]	G	
	C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T	
		CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C	
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A	
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G	
	A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T	
		ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C	
		ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A	
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G	
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T	
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C	
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A	
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G	

Figure 1.20: Genetic Code

In the genetic code table (Figure 1.20), there are four important codons to be noted here, i.e. ATG, TAA, TAG, and TGA. Please note that in RNA, T is exchanged with U. What is so special about these 4 codons? ATG is the start codon, usually translation start from Methionine, therefore it works some sort like a start signal. Whereas TAA, TAG, and TGA are the stop codon. Whenever, ribosomes (the protein synthesizing machines) come to the stop codons, the ribosomes will dissociate from the mRNA and the translation process terminates.

1.6.1.4 More on Gene Structure

Now let's take a look at the structure of a gene (see Figure 1.21).



Figure 1.21: Gene Structure

A gene consists of three regions: the 5' untranslated region, the coding region, and the 3' untranslated region. The coding region contains the codons for protein. It is also called “open reading frame”. Its length is a multiple of 3 since each codon consists of three nucleotides. The coding region usually begins with a start codon, and must end with an end codon, and the rest of its codons are not an end codon. The 5' untranslated region, coding region and 3' untranslated region together are also called the “mRNA transcript”, because it is exactly what the mRNA copied from the DNA. And as we have mentioned before, there are some regions in the gene that are “useless”, i.e., they will not be translated into protein. These “useless” regions are just the two “untranslated” regions here. Finally before the 5' untranslated region, we have the regulatory region (also called the promoter) which regulates the transcription process. The promoter is in fact a DNA molecule to which RNA polymerase binds, initiating the transcription of mRNA.

1.6.1.5 The translation process

Now we can discuss the translation process in details. The translation process is handled by a molecular complex ribosome which consists of both proteins and ribosomal RNA (rRNA). First, the ribosome reads mRNA from 5' to 3'. The translation starts around the start codon (translation start site). Then, with the help of transfer RNA (tRNA, a class of RNA molecules that transport amino acids to ribosome for incorporation into a polypeptide undergoing synthesis.), each codon is translated to an amino acid. Finally the translation stops once ribosome reads the stop codon (translation stop site).

1.6.1.6 More on tRNA

We have said that the translation from codon to amino acid is with the help of the transfer RNA, or “tRNA”. Totally there are 61 different tRNAs, and each corresponds to a non-terminated codon in the genetic code table. Each tRNA folds to form a cloverleaf-shaped structure. On one side it holds an anticodon (A sequence of three adjacent nucleotides in tRNA designating a specific amino acid that binds to a corresponding codon in mRNA during protein synthesis), and on the other side it holds the appropriate amino acid. The structure of the tRNA is shown in Figure 1.22.

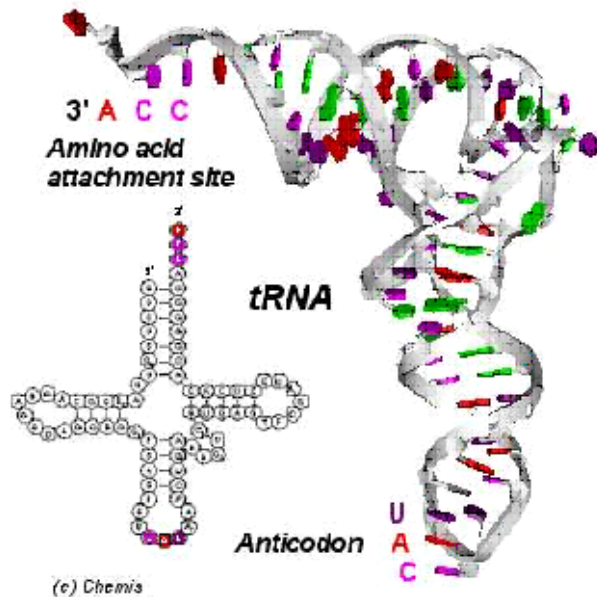


Figure 1.22: tRNA Structure

1.6.2 Central Dogma for Eucaryotes

After discussing the Central Dogma for the Prokaryotes, we move to the Eukaryotes. In Eukaryotes, transcription is done within the nucleus. The resultant mRNA is then carried out from the nucleus and the translation occurs outside the nucleus. See Figure 1.23.

1.6.2.1 Introns and exons

The coding region of an Eukaryote's gene is different from that of a Prokaryote. For Eukaryotes, each gene contains Introns and Exons. Intron is a segment of gene situated between exons. It is not responsible for the coding of protein. So the Introns will be ultimately spliced out of the mRNA. And Exon is a nucleotide sequence in DNA that carries the code for the final mRNA molecule and thus defines the amino acid sequence during protein synthesis. The process of removing the introns for the mRNA sequence is called RNA splicing. This process is done with the help of spliceosomes. Though the Introns seem "useless", it is quite amazing that in Eukaryotes, each gene can have many Introns, and each Intron may have thousands of bases. Introns in eukaryotic genes normally satisfies the GT-AG rule, that is intron begins with GT and ends with AG. Introns can be very long. An extreme example is the gene associated with the disease cystic fibrosis

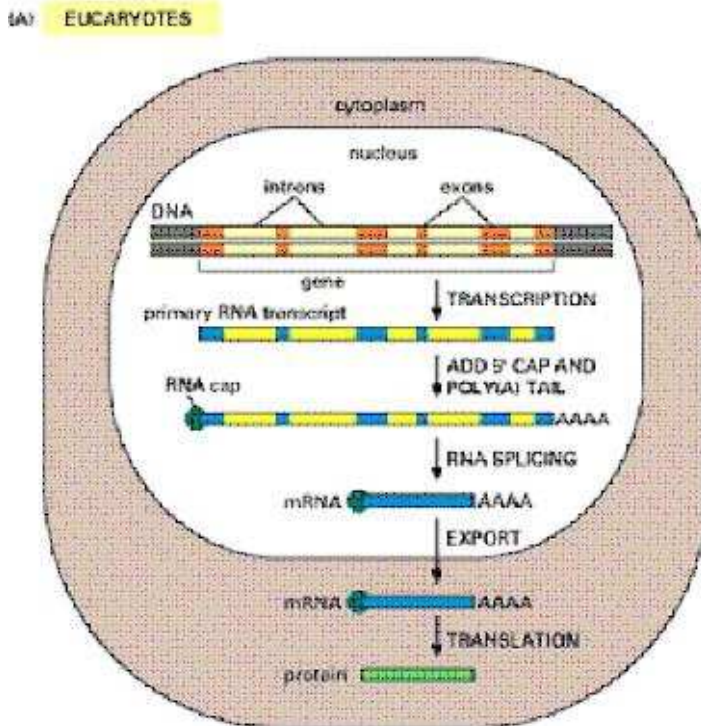


Figure 1.23: Central Dogma for Eukaryotes

in humans, i.e. it has 24 introns of total length approximately 1 MegaBases. Whereas the total length of its exons is only 1 kilobases.

Finally, note that in different tissues or different conditions, the introns which will be removed are not the same. Hence, one gene can be transcribed into many different mRNA sequences.

1.6.2.2 Transcription (Eukaryotes)

Now we discuss the transcription process of Eukaryotes. Firstly a pre-mRNA is produced which contains both Introns and Exons. Then the 5' cap and poly-A tail are added to the pre-mRNA. After that, the RNA splicing removes the Introns and mRNA is produced. Finally the mRNAs are transported out of the nucleus and it is translated to a protein using exactly the same process as described in Section 6.1.5.

1.7 Basic Biotechnological Tools

A vast range of technological tools have been developed to facilitate the scientists to study DNA in a more efficient manner. Basic tools help to cut and break DNA (using Restriction Enzymes, or Shotgun Method), to duplicate DNA frag-

ments (using Cloning, or PCR), and to measure the length of DNA (using Gel Electrophoresis). Each of these tools is examined in this section.

1.7.1 Restriction enzymes

Restriction enzymes or restriction endonuclease is a class of bacterial enzymes. They are DNA-cutting enzymes, which recognize certain point, called restriction site, in the double-stranded DNA with a specific pattern and break the phosphodiester bonds between the nucleotides. Such process is called digestion. Naturally, restriction enzymes are found and isolated from various bacterial species, which are used to break foreign DNA to avoid infection or disable the function of the foreign DNA.

Each type of restriction enzyme seeks out a specific DNA sequence, which is palindromic and usually 4 to 8 bp long, and precisely cuts it in one place. For instance, the enzyme shown here, EcoRI, cuts the sequence GAATTC (see Figure 1.24), cleaving between the G and the A. If GAATTC is a palindrome, GAATTC will be its own reverse complement. Depending on the types of restriction enzymes used, DNA fragments with either blunt ends or sticky ends can be produced. For EcoRI, a sticky end is formed after the digestion.



Figure 1.24: Digestion by EcoRI

Restriction enzymes are used in cloning application(see Section 7.3), where the target DNA and the vector are cut with the same restriction enzyme. It is also used to cleave DNA molecules, producing specific fragments, which can then be subsequently fractionated and determined.

1.7.2 Shotgun method

Shotgun method involves randomly chopping a DNA fragment, with no foreknowledge of where to cut. Basic idea is to apply high vibration(eg. sonication) to a solution which has a large amount of purified DNA, so that each DNA molecule is broken randomly into small fragments. Shotgun method has been applied to sequence the whole genome.

1.7.3 Cloning

Given a piece of DNA X, the process of duplicating it into many pieces is called cloning. The basic steps involve:

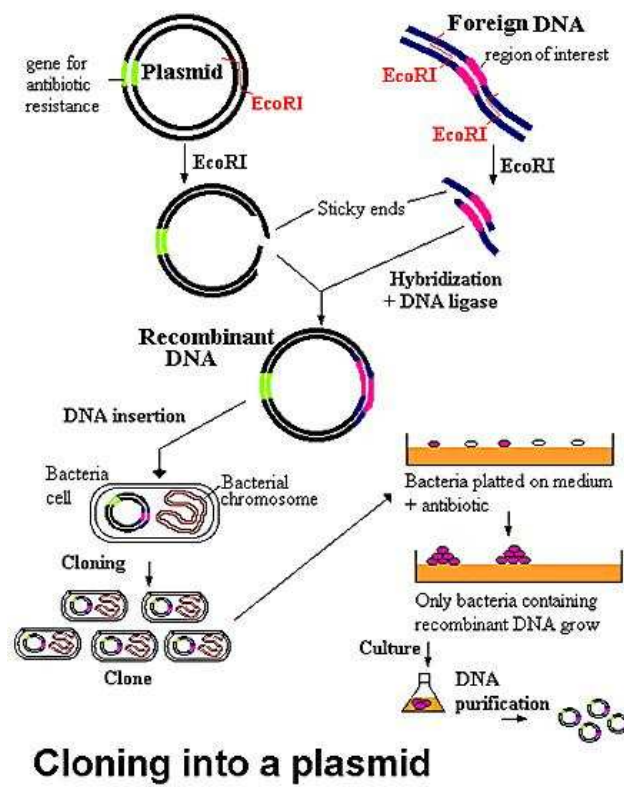


Figure 1.25: Cloning of foreign DNA into plasmid

1. Insert X into a plasmid vector with antibiotic-resistance gene and a recombinant DNA molecule is formed. Plasmids and DNA fragments must have compatible RE ends for ligation by T4 DNA ligase. A linear product of DNA and the linearized plasmid is firstly formed, followed by the joining of the opposite ends to form a circular product.
2. Insert the recombinant into the host cell (usually, *E. coli*). This makes use of a chemical based transformed method, where the bacterial cells are made “competent” to take up foreign DNA by treating with calcium ions. After the recombinant DNA molecules are mixed with the bacteria cells, a brief heat shock is applied to facilitate uptake of DNA.
3. Grow the host cells in the presence of antibiotic. Note that only cells with antibiotic-resistance gene can grow. Note that when we duplicate the host cell, X is also duplicated.
4. Select those cells contain both the antibiotic-resistance genes and the foreign DNA X. Some cells only contains plasmid vector but without the foreign DNA due to unsuccessful ligation in step 1. The cells with foreign DNA X can be correctly selected by the α -complementation of beta-galactosidase, in which the correct colony will show blue colour.
5. Kill them and extract X.

Cloning is a time-consuming process normally requiring several days, whereas we will learn about a much quicker method known as PCR in the next section. Both cloning and PCR multiply the available amount of DNA in order to have enough DNA for many experiments.

1.7.4 PCR

PCR is an acronym which stands for polymerase chain reaction. What is a polymerase? A polymerase is a naturally occurring enzyme, a biological macromolecule that catalyzes the formation and repair of DNA. The accurate replication of all living matter depends on this activity. In the 1980s, Kary Mullis at Cetus Corporation conceived of a way to start and stop a polymerase's action at specific points along a single strand of DNA. What is the chain reaction? Mullis also realized that by harnessing this component of molecular reproduction technology, the target DNA could be exponentially amplified.

The PCR technique is basically a primer extension reaction for amplifying specific nucleic acids *in vitro*. The synthesis of DNA begins with binding of primers to the DNA template and single nucleotides (dATP, dCTP, dGTP & dTTP) are added to the 3' end of the growing DNA molecule. The use of a thermostable polymerase which tolerates the high temperature of up to 95°C

in the denaturing step allows the dissociation of newly formed complementary DNA and subsequent annealing or hybridization of primers to the target sequence with minimal loss of enzymatic activity. Inputs for PCR includes: (1) Two oligonucleotides are synthesized, each complementary to the two ends of the DNA fragments to be amplified. They are used as primers. (2) Thermostable DNA polymerase TaqI.

PCR consists of repeating a cycle with three phases 25-30 times. Each cycle takes about 5 minutes.

Phase 1: The denaturing step of separating double stranded DNA by heat;

Phase 2: Cool; add synthesis primers to anneal to the denatured DNA;

Phase 3: Add DNA polymerase TaqI to catalyze 5' to 3' DNA synthesis.

After the last cycle, Phase 3 is kept for a longer time at about 10 minutes to ensure that DNA synthesis for all strands are complete. Then, only the flanked region by the primers has been amplified exponentially, while the other regions are not. For a comprehensive picture of how PCR works, please see Figure 1.26.

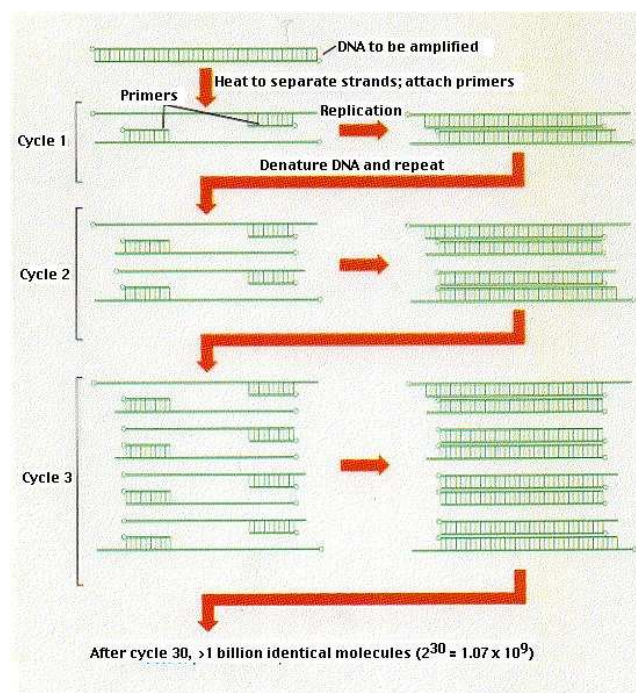


Figure 1.26: Cycles in Polymerase Chain Reaction

PCR method is used to amplify DNA segments to the point where it can be readily isolated for use. When scientists succeeded in making the polymerase chain reaction perform as desired in a reliable fashion, they had a powerful technique for providing unlimited quantities of the precise genetic material for doing experiment. Some examples of PCR applications are: (1) Clone DNA fragments

from mummies; (2) Detection of viral infections.

1.7.5 Gel electrophoresis

Gel electrophoresis, developed by Frederick Sanger in 1977, is a method that separates macromolecules—either nucleic acids or proteins—on the basis of size, electric charge, and other physical properties.

A gel is a colloid in a solid form. The term electrophoresis describes the migration of charged particle under the influence of an electric field. Electro refers to the energy of electricity. Phoresis, from the Greek verb phoros, means “to carry across.” Thus, gel electrophoresis refers to the technique in which molecules are forced across a span of gel, driven by an electrical current. Activated electrodes at either end of the gel provide the driving force. A molecule’s properties determine how rapidly an electric field can move the molecule through a gelatinous medium.

Organic molecules such as DNA are charged. DNA is negatively charged due to the high phosphate residues in their backbone. A gel is prepared which will act as a support for separation of the fragments of DNA. Holes are created in the gel. These will serve as a reservoir to hold the DNA solution. DNA solutions (mixtures of different sizes of DNA fragments) are loaded in a well in the gel. During electrophoresis, DNA migrates towards the positive electrode. The greater the charge on the DNA molecule, the faster it migrates. However, the movement is retarded by the gel matrix, which acts as a sieve for DNA molecules. Large molecules have difficulty getting through the holes in the matrix. Small molecules move easily through the holes. Because of this, large fragments will lag behind small fragments as DNAs migrate through the gel. As the process continues, the separation between the larger and smaller fragments increases. The mixture is separated into bands, each containing DNA molecules of the same length.

An application of gel electrophoresis is to reconstruct DNA sequence of length 500-800 within a few hours. The idea is to, firstly, generate all sequences end with A. Then we can use gel electrophoresis to separate the sequences end with A into different bands. Such information tells us the positions of A’s in the sequence. Similarly, we can process for C, G, and T accordingly.

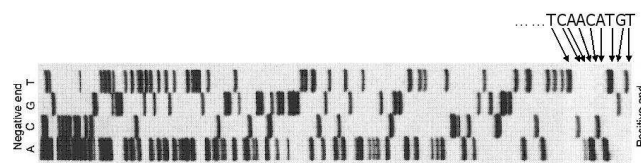


Figure 1.27: Autoradiograph of a Dideoxy Sequencing Gel

All the four group of fragments: A, C, G and T are placed at the negative end of the gel. During electrophoresis, the fragments move towards the positive end. The unknown DNA sequence is reconstructed from the relative distances of

the fragments as shown in Figure 1.27. The letters A, C, G and T at the negative end refers to ddATP, ddCTP, ddGTP and ddTTP reaction mixtures respectively. Usually, the obtained sequence is verified by carrying out the same sequencing method on the complementary strand.

The generation of fragments ending in a particular base can be achieved through two methods, i.e. Maxam-Gilbert Sequencing or Sanger Method. In Maxam-Gilbert Sequencing, DNA samples are divided into four aliquots and four different chemical reactions are used to cleave the DNA at a particular base (A, C, G or T) or base type (pyrimidine or purine). In Sanger Method, DNA chains of varying lengths are synthesized by enzymes in four different reactions. Each reaction will produce DNA ending in a particular base.

DNA sequencing can be automated by using laser to detect the separated products in real time during gel electrophoresis. The four bases are labelled with different fluorescence and they are placed in a single lane. In this way, many DNA samples can be sequenced at 1 time. This can sequence a total of 900 bases in length and achieve a 98% accuracy well beyond 900 bases. An example of electropherogram is shown in Figure 1.28 below.

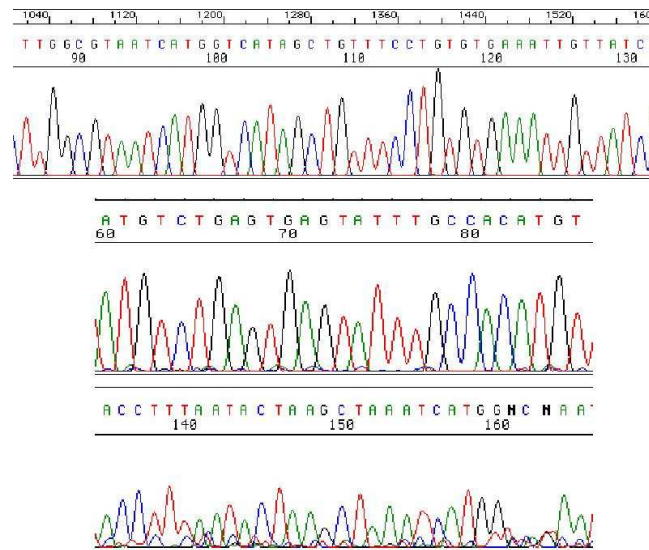


Figure 1.28: Electropherogram from Automated DNA Sequencing using fluorescent dyes

1.7.6 Hybridization

Routinely, biologists need to find a DNA fragment containing a particular DNA subsequence among thousands of DNA fragments. The above problem can be solved through hybridization in the following steps:

1. Suppose we need to find a DNA fragments which contains ACCGAT

2. Create probe which are inversely complementary to ACCGAT
3. Mix the probes with the DNA fragments
4. Due to the hybridization rule (A=T, C=G), DNA fragments which contain ACCGAT will hybridize with the probes

1.7.6.1 DNA Array

The idea of hybridization leads to the evolution of DNA array technology, which enable the researchers to perform experiment on a set of genes or even the whole genome. This completely changes the routine of one gene in one experiment and make easier for researchers to obtain whole picture on the particular experiment.

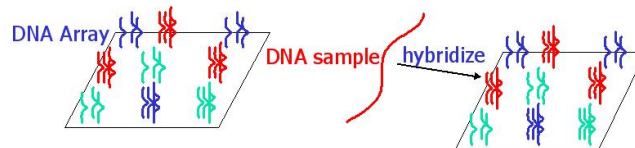


Figure 1.29: Hybridization of DNA samples and Probes

DNA array is an orderly arrangement of thousands of spots, each containing many copies of the same DNA fragment. When the array is exposed to the target solution, DNA fragments in both array and target solution will match based on hybridisation rule: A matches to T while C matches to G through hydrogen bonds. Such idea allows us to do thousands of experiments at the same time.

1.7.6.2 Application of DNA Arrays

DNA arrays can be used for sequencing using its hybridization method. The basic idea is that if all the subsequences of the target sequence are presented in the DNA array chip, after hybridization, the target sequence can be constructed from overlapping of oligonucleotides by algorithms from the spectrum of intensities generated at every spots. However, due to the presence of repetitive DNA that obstructs reconstruction, the method is unlikely to be used in sequencing complex genomes, but useful for short non-repetitive DNA fragments. The fabrication of DNA chip can be done by combining the photolithographic method and combinatorial synthesis of oligonucleotides on the surface of glass chips.

DNA array can also help researchers to investigate the expression profile of a cell. Through spotting the complementary sequence of genes on the DNA chip, the activities within a cell can be monitored by observing the expression levels of these genes at different conditions or time points. Due to hybridization, we can also measure the concentrations of different mRNAs within a cell.

There are many other applications where DNA arrays can be put to good use.

1.7.6.3 More Advanced Application – Mass Spectrometry

Mass spectrometry (MS) is one type of instruments that is used for analyzing biomolecules, particularly proteins and peptides. MS is very important as a tool in proteomics field. MS offer three types of analyses in Proteomics analysis. First, MS can provide highly accurate protein mass measurements compared to the traditional gel electrophoresis (in this case, SDS-PAGE) which is not sufficiently sensitive. Secondly, MS can also provide accurate mass measurements of peptides from proteolytic digests. In contrast to whole protein mass measurements, peptide mass measurements can be done with higher sensitivity and mass accuracy. The data from this peptide mass can be searched directly against the databases, and frequently to obtain definitive identification of the target proteins. Finally, MS can also provide the sequence of peptides obtained from proteolytic digests. Indeed, MS is now considered the state-of-the-art in peptide-sequence analysis. MS sequence data provide the most powerful and unambiguous approach to protein identification.

1.8 List of Bioinformatics Problems

1.8.1 Biological Data Searching

There has been an explosion in the amount of biological data, resulting from various scientific efforts such as Human Genome Project. As such, it is important for the biologist to be able to retrieve the relevant information from such a huge pool of resources. The difficulties faced in the computational approach is that the data are too enormous. Consider human genome alone, there are 3G bases! There are also many mutations occur in the genomes, impeding the process of finding homologies, motifs and etc. Thus, it is important to have efficient algorithms for locating approximate matches.

1.8.2 Gene / Promoter finding

It is expected to have 30k-35k genes in human genome. However, we only found thousands of genes up to now. It is necessary to have some good methods to find the remaining genes.

1.8.3 Cis-regulatory DNA

Cis-regulatory DNAs controls whether genes should be expressed or not. Cis-regulatory elements may locate in promoter region, intron or exon. They may be extended over much longer stretches of DNA involving thousands of nucleotides.

1.8.4 Gene Network

Inside a cell is a complex system where there are extensive intra-cell or cell-cell communication and physical configuration in three-dimensional space. The expression of one gene depends on the expression of another gene. Such interactions can be represented using gene network. Understanding such network helps to identify the association between genes and human diseases. From large scale gene expression data, either in time series or steady-state data, the gene regulatory network can be inferred.

1.8.5 Protein / RNA Structure Prediction

The three dimensional structure of Protein/RNA is essential to its functionality. The shapes, properties of the key functional areas of the structures carry out the protein's biological functions. The structure must also be stable in the normal environment of the protein. Therefore, it is very important to have some ways to predict the structure if a protein/RNA is given its sequence. This problem is important and it is always considered as a *grand challenge* problem in bioinformatics.

1.8.6 Evolutionary Tree Reconstruction

Recall that protein/RNA/DNA can mutate. Through the evolutionary tree or phylogenetic tree, the evolutionary relationship among a set of proteins/ RNAs/ DNAs can be studied. The different species are represented by nodes and the ancestor-descendent relationships are represented by edges. However, caution must be exercised as the sequences may be subjected to different mutation rate and differential selection. Some of the methods for building the evolutionary trees are distance matrix methods, maximum parsimony methods and maximum likelihood methods.